

医療とコンピュータ

Vol.9 No.4

株式会社日本電子出版

<http://www.epj.co.jp/medcom/>



ソフトウェア活用法： 統計解析

陶山 昭彦 （鳥取大学医学部衛生学）

統計処理における超解析法

「超」と聞くとコンピュータの世界では、ひと昔前で「超高速」演算 (Super Computing) であり、現在だと並列コンピュータでの並列Fortranによる「超並列」演算 (Hi-Performance Computing) といった計算資源の効率化がある。一方で「超整理法」、「超勉強法」といった知的生産技術の効率化が今どきのキーワードになってきている。読者にとっては「超very good (最適) な」統計解析法、そしてその解析法を包含する統計解析ソフトの答えを希求する反映としての「超(統計)解析」法であるかもしれない。そのどれもが超むずかしい問いかけなのだが...

「超」解析法 その1

研究仮説の明確化と検証手段の最適化

わが国における医薬品承認の問題点がうきばりにされてくる中、医薬品承認の統計的側面の重要性が指摘されるようになってきた(文献1)。新GCP実施に伴い臨床医師も薬効評価に関して、統計学的知識を避けて通れない状況がみえてきている気がしてならない。疫学領域でも研究デザインの企画が特に強調されている(文献2)。薬効、毒性について検証的な側面から、厳格に無作為割付試験が実施され検出力が制御された統計解析法に従って検討される領域では、「証明すべき仮説は明確に、データ解析によって検証できる形に述べられて」おり「妥当な解析計画」にもとづいて「解析結果の解釈を適切」(文献3、大橋)に行うこと

が重要である。特に関心を払いたいのは「より探索的な相においてはモデルを想定したmodel-basedな解析により効果的に情報を縮約し、検証的な相においては前提を必要としないdesign-basedな解析を重視する」(大橋、同文献)と指摘されている。臨床試験第I相～第IV相に関して検証仮説の意図を明確にもち、検証手段としての統計解析法を最適化するためには、系統的な生物統計学の基盤の上に医学研究の現場で訓練された生物統計学者の参画が極めて重要である。また医師の立場においても実例によって科学的事実に対する仮説設定と検証のための解析力の訓練をつみあげていくことが大切と思っている(文献4)。

ソフトウェアを利用した、検出力にもとづく標本数算出 - SamplePower -

統計学の教科書の巻末にしばしば掲載されている表は重要な統計表であるにもかかわらずあまり省みられることが少ない。National Institute of Mental HealthとSPSS社で開発された「SamplePower」は比較的良好に知られた解析手法において、検出力から標本数を求めることができるソフトである。図1の例の場合、2群の平均値の比較において、平均値がそれぞれ21.3と21.2、SDが1.0でalphaが0.05で両側検定の場合、統計的検出力（power）が0.8となるために必要な症例数が「Find N for required power of 80%」コマンドで1571例と求められる。評価版が30日の使用期限つきでSPSS社のURL（<http://www.spss.com/software/samplepower/>）からダウンロードできる。必要症例数を求めるためのフリー

ウェアは後述のStatLibにも登録されているが、SamplePowerは統計的検出力と必要症例数の関係を理解する上でも有用なデザイン設定のための支援ソフトであり、是非試してみることをすすめたい。

一元配置分散分析とGLM

SPSSがVer.7.5になって、一般線形モデル(General Linear Model)が扱えるようになった。StatViewでは、別売のSuperANOVAでカバーしている部分であり、SASではGLMプロシジャが相当する。実験計画法には1元配置、2元配置などの手法や測定の繰り返し数は同じ場合（釣り合い型）と異なる場合（不釣り合い型）など複雑な状況に対応して解析しなければならないことがほとんどだ。一般線形モデル（GLM）はこれらをひとつのモデルを使って柔軟に解析できる特徴がある。このことを理解するための

解説は他書（文献5）にゆずるが、GLMにおいて経時的測定データ（Repeated measurement data）が扱える。（図2）SPSS Ver.7.5ではこのように医学的研究実験に必要な解析手法であるGLMが組み込まれたことが大きな特徴といえるだろう。

JMPによる解析

今後、StatViewはJMPと統合される（コラム参照）とSAS社からアナウンスされていることから、JMPユーザーが少ない医療関係者にJMPによる解析の雰囲気は知っておくべきだろう。

JMPの大きな特徴は選択すべき解析手法のメニューが表面的にはあらわれておらず、変数を選択した段階で自動的に解析法が決定する。（図3）

はじめてJMPを使用した時には非常に面くらうだろうが、慣れてみるとたいして違和感はない。Cox比例ハザードモデルやロジスティック回帰モデルにstepwiseオプショ

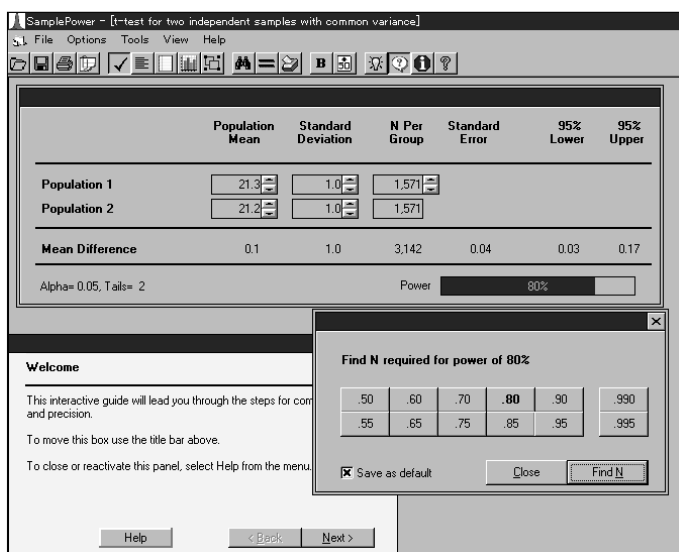


図1



図2

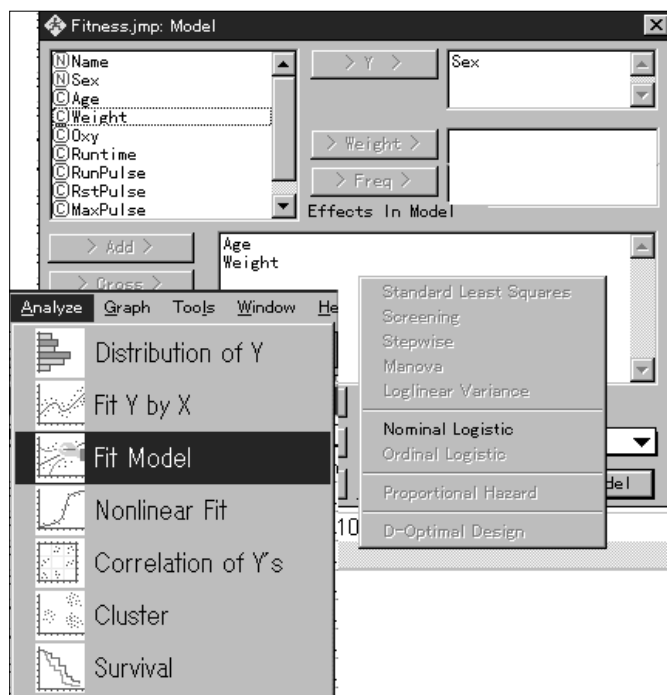


図3



図4

データ量からも計算資源の能力が不足しているということは考えにくい。といっても、医療関係者は圧倒的にMacユーザーと思われ、G3マシンの公表によりマシンの乗り換えを検討している読者も多いと思う。

統計解析は超高速とまでいなくてもマシンパワーが要求される処理に属する。この意味からいえば、統計解析についてはパーソナルなレベルでも計算資源を効率的に生かしたほうが有利である。

ンがないのが残念だ。自動的に解析法が決定されるという意味では「超」解析ソフトといえそうだ。JMPの詳細な操作法については他書を参考にして頂きたい。(文献6)

2 超解析法 その2 - 計算資源の効率化 -

UltraSPARCやAlphaのようなRISC CPU搭載のワークステーションには手がでなくても、PentiumIIやPowerPC G3搭載の今どきのパソコン環境下にあっては、計算量や

コラム StatViewが、Mac版SPSSが消える！？

来るものが来たという感じを受ける。この執筆依頼をうけて久しぶりにAbacus社のホームページをのぞくと、SAS社が昨年の9月づけでAbacus社からStatViewを買いとったとある。(図4) Abacus社のCEOのFeldman氏の言も掲載されていることから裏付けられよう。StatViewと並んで、理系研究者から根強い支持を受けているSAS社の統計ソフトJMPの開発者でありSAS者の創立者の一人であって統計学者でもあるJohn Sall氏によるとStatViewはJMPと統合して新しくリリースする、とのことである。SPSS社はSYSTATなどよく知られた統計ソフトなどを買収してはいるが、販売元がSPSS社に移行しただけでソフトウェアの環境の変更があったわけではないようだ。着々とサイエンスソフトウェアの統廃合が進んできていたのだが、StatViewへのJohn SallのいれこみはWebページでも見て取れ、長きにわたってStatView

取得を目論んでいたようであり、SASシステムに組み込まれている多くの統計学的手法が洗練されてJMPに組み込まれており、StatViewのGUIインターフェースの操作性が良いところをうまくとりこんでくれるとJMPとStatViewの両刀づかいのユーザーとしてはうれしいところではある。

しかし、以前StatViewがVersion 4.*に移行したときに医療関係者が大混乱しただけに、次に出現するだろうStatViewとJMPの合体化身がどのようなものなのか想像がつかない。

もうひとつMac版サイエンスソフトが次々の開発中止となってきた現状がある。もっとも淋しいアナウンスはSPSS社からWin版ではVer.7.5へバージョンアップされたのに対して、Mac版はVer.7.5は開発しないということだ。これにより、繰り返し測定などに適したGLMコマンドはMacユーザーは恩恵が蒙れない。Mac雑誌の日本のMacUSERも休刊となり、次第に外堀がうめられてきた感がある。

Win95/NTとMacの併存統計解析環境

グラフィックとDTP領域でMacが優位だったが、PhotoshopやIllustratorやCanvasなどWin版ソフトが次第に最適チューニングされてきている。高速画像処理のCG領域はもうWinNTが先行してしまっている。コラムで述べたように、主軸ソフトが欠落したMacは裸の王様になってしまう危険性がある。統計解析においては、もはやWinマシンに委ねるべきである。Winの場合、ショップが既存パーツを最適化したマシンの方がシステム設定時に困りにくい。AMD K5搭載したものにいたっては10万を切っていて、最速のPII搭載のマザーボードにのりかえても10万少しで変更できる。StatViewの場合、計算結果が解析ワークシートに保留されているため、探索的に解析を行っているため、ワークシートにリンクしている変数に対して変更がおきると、データ数に比例して莫大な再計算の待ち時間が生じる。同じくStatViewで処理時間が長いものに外部テキストファイルの読み込みがある。この2つだけに限っても、Winマシン上で処理させた経験から、一昔前のPentium 133MHzでもMacと比較して著しく早い。LANカードやSCSIカードをつけても3万程度なので、Hub経由でMacとWinを接続しよう。MacとWinのファイル共有は安上がりなPC-MacLANがCopsTalkを導入してAppleTalkに

より共有しておくことにより互いにデスクトップ上からファイル操作が可能になる。(文献7)これでMacで作成した統計解析ファイルをWinへ逐次移行していくことが可能になる。同じ統計ソフトでもMac版とWin版は作成されたファイルはダイレクトに読み込めることが少ない。執筆にあたってSPSS、JMP、StatView、StatisticaのMac版ファイルをWinフォルダへコピーしたのち、各ソフトでファイルを開くことができるかどうか調べたが、全く問題なく互換があったのはSPSSのみであった。

Win環境でのファイルインターフェースにはMicrosoft社のODBC(Open Data Base Conector)を利用する機能が設定されていることが多い。

ODBCを利用すると、ACCESSで構築しているデータベースを統計解析用の作業用データファイルを作成しなくても、直接データベースファイルに接続することができる。また、SPSSでは利用権限が設定されたデータへのアクセスやデータベース内部のリレーションシップもとりにくくすることが可能となるので、データベースによるデータ管理は独立して維持できる。Win環境ではAccessを軸にデータサーバーとパーソナルクライアントにより簡単に医局内データ共有ができ、ODBCにより解析ソフトからファイル接続できるよう方向づけておくことが望ましい。

Macでしかできない処理も極めて多いので、Winマシンをしばらくは解

析専用のLANに接続された装置とみたとしても、各種周辺装置よりは安い投資になるはずである。

Webからの統計解析 - 超高速コンピュータの力をひきだす -

統計解析において計算資源が効率化できるアプローチとしてWebインターフェースを利用することがある。DNAの相同性検索は不定長のパターンマッチング処理で恐ろしく計算資源を消費する。国立遺伝研究所はUNIXやスーパーコンピュータでのDNA解析をコマンド投入モードだけでなくWWWを使ってブラウザから可能にしている。統計解析についても少しずつブラウザから実行しブラウザ上で解析結果が得られるようなインターフェースが現れてきた。ようやくリリースされたばかりだが、SAS/InterNetはSASの操作プラットフォームをブラウザ上に実現できるプロダクトと説明されている。(<http://www.sas.com/japan/m/special/webdemo.html>) 統計ソフトをWebから操作するという試みではJMP Macintoshの場合、AppleScriptによりJMPが制御できる部分が公開されているため、CGIによりMacをJMPサーバーとみたとて、クライアントのブラウザからJMPの統計解析機能を利用することもできる(文献8)。

この場合、JMPの解析エンジンが利用できるのが簡単であるが、Java



で解析ルーチンまでつくりこめば、自作統計解析ソフト on Webということで公開でき、インターネットから、そのソフトが公開されているマシンの計算資源を有効に利用できる (<http://aoki2.si.gunma-u.ac.jp/BlackBox/BlackBox.html>)。これから期待できるかもしれない統計解析の環境であろう。

統計解析ソフトが利用できる 情報処理センター

国立大型計算機共同利用センターとして北海道大、東北大、東京大、名古屋大、京都大、大阪大、九州大があげられる。SASやSPSSが多くの場合利用できるが、多くのセンターがメインフレーム上で統計解析ソフトを運用している中で、九州大のセンターはSASのサービスをunixサーバーで提供している。 (<http://www.cc.kyushu-u.ac.jp/users/545.html>) センターのメインフレームマシンは通常24時間運転されていないことが多いが、unix上のSASサーバーは利用資格さえあれば、インターネット接続で利用可能である。X 端末エミュレーションでもSASが操作できるので、動作はかなりにぶいが統計解析機能とGUI機能を生かしてSASサーバーの資源を利用できる。ただし、SASのX上の操作はCPU資源を多く消費するようなので利用請求金額に注意しておく必要がある。

3 超解析法 その3 - 知的生産技術の効率化 -

インターネット情報源の利用

医師にとっては統計学領域は非常に越えがたい敷居であるかのようである。これを補完してくれそうなのが、国内国外のホームページの情報源である。熱心な統計学者が丁寧な解説をしていたり、有用なソフトに関する情報を掲載していたりするので、まめにチェックしてみる価値は大いにある。

海外のインデックスサイトとしてはVirtualLibraryで統計学 (<http://www.stat.ufl.edu/vlib/statistics.html>) や生物統計学 (<http://www.biostat.washington.edu/Xvlib/>) をキーサイトにするとよいだろう。

国内では今のところよく整理されたインデックスサイトはなく、各研究者が整理したリンクインデックスを参考にしていくのがよい。国内研究室のエントリーとしては岡山大学統計学教室 (<http://stat1.stat.ems.okayama-u.ac.jp/>) や東京理科大学仁木研究 (<http://nikii.ms.kagu.sut.ac.jp/indexj.html>)、東京大学廣津研究室 (<http://www.stat.t.u-tokyo.ac.jp/toukeilink-j.html>)、群馬大学青木研究室 (<http://www.si.gunma-u.ac.jp/aoki/Yogoshu/words.html>) などの統計学関連リンクなどが非常に参考になる。このうち青木先生のホームページでは統計学用語

集 (<http://www.si.gunma-u.ac.jp/aoki/FlowChart/Tutorial.html>) が非常によく整理されている。

ソフトウェアサイトとしては、先のVirtual Libraryから枝葉をたどっていけば過去に作成された価値あるものが蓄積されている。多くの統計学者のページからリンクされているStatLib (<http://lib.stat.cmu.edu/>) は統計ソフトウェアのインデックスサイトとして貴重である。

解析出力の管理法

とりだてて超出力管理法というほどのことでもないのだが、生存率解析などにあたって、生存率曲線の比較やノンパラメトリックな生存率の差の検定、Coxモデルによる単変量解析や多変量解析などを数項目にわたって検討していた場合、上手に出力を管理していないと、どの解析がどの出力に対応するのが区別できなくなり、同じ解析を繰り返さなければならないはめになってしまうことも多い。

・出力ナビゲータ - SPSSの新機能 -
SPSSは以前のバージョンでは、解析結果がlogファイルに追記されている、またファイルには小さなサブ出力ごとに改頁コードがはいっており、非常に扱いにくかった。そのままの設定で出力紙枚数が膨大になった上、2、3行ほどの印字出力だけでも改頁していた。この点、JMPはひとつの処理ごとに出力をつくるため印字出力する場合は制御し

やすい。StatViewは前述したように出力結果のワークシート上に変数が保留されたままになっているため、完全にテキストファイルとしてファイルにしておかないとわずかな変数変更により再計算に悩まされる。前述したように、単変量や多変量などを網羅的に探索して検討する場合などは、むしろSPSSのように出力logに追記されたほうが、解析ストーリーがつかめていいのだが、logファイルが非常に単調なために、カテゴリー変数を生成して中途半端な変数名で出力して生成履歴をドキュメント化しない場合、非常にストーリーが把握しにくくなる。今回Ver.7.5になって、出力ナビゲーターという機能が付加され、処理ごとにExplorerのような環境でタイトル、テキスト結果、グラフ結果が階層的に形成される。(図5) 図で示した共変量(性)と共変量(性、肥満度)の出力はブロックごとで順序が入れかえることが可能であり、必要な解析ストーリーに沿って出力を編集できるきわめて

自由度が高い編集機能である。

・PDFファイルとして出力する

SPSSのような出力ナビゲーション機能のかわりに、テキストとグラフを同時にファイル出力して、処理ごとに追記する方法がある。これはAdobe Acrobatを導入して、印刷ドライバにAcrobat PDF Writerを利用することで、処理ごとに出力をPDF化できる。Acrobat ExchangeのPDF編集機能を利用することによって、それぞれのPDF出力を合体させ順序だてて出力をながめながら解析の検討を行うというものである。やや手間がかかるといえはかかるが、出力結果を効率的に管理しておくためにはPDFファイル化は管理方法が簡単であるため推奨したい。

研究室の机が資料で山積みになっているように、パソコンの中もMacといわず、Windowsといわずもう貴重なファイルなのかゴミなのか判別不能な状態である。統計解析は研究仮説から腰をすえてかからねば

ならない仕事のひとつであり、これらの「超」(低級?)解析法が読者の何らかの役にたてれば幸いである。

参考文献

- 1) 吉村功他：臨床試験における統計学上の諸問題、臨床医薬、1993、Vol.5
- 2) 陶山昭彦(共訳)：データ処理と解析の方法を計画する、医学的研究のデザイン(木原 正博監訳)、医学書院MYW、東京、1997
- 3) 大橋靖雄：薬効評価に必要な実験計画法と統計手法、臨床薬理学、医学書院、東京、1992
- 4) 陶山昭彦他：研究仮説に対する統計ソフトによる検定アプローチ、別冊医学のあゆみ「研究室でのパソコン活用術」、医歯薬出版、東京、1996
- 5) 高橋行雄他：SASによる実験データの解析、東京大学出版会、東京、1989
- 6) 陶山昭彦、祝部大輔：医学統計解析テクニック、BNN、東京、1993
- 7) 陶山昭彦：パーソナルFTPのすすめ、The BASIC No.165、p.88-92、1997
- 8) 陶山昭彦、井上仁：JMP3.0を利用したWWW統計解析ホームページ、SUGI-J'96/IDS Conference '96論文集、459-466、1996

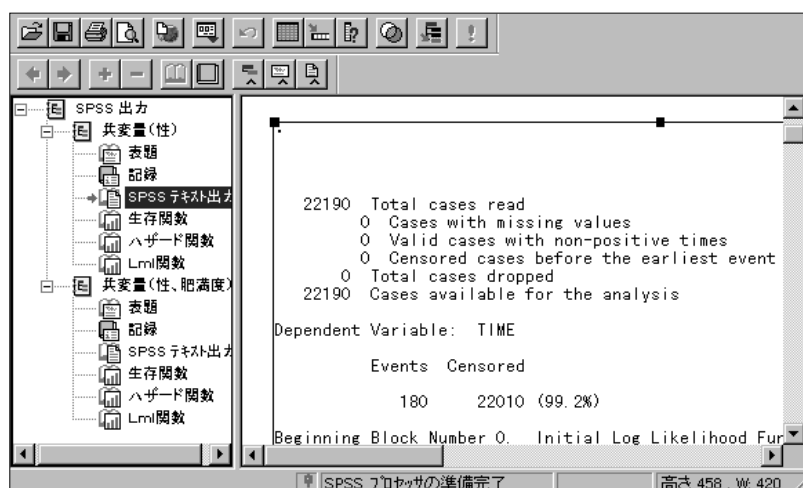


図5